ISSN: 2664-9640

Groundwater level prediction method based on LSTM

Shaofa Zhou, Zhongyi Liu and Li Lu*

School of Yangtze University, Jingzhou 434000, Hubei, China

*Corresponding Author

Abstract

Water is an essential substance for human existence, and groundwater, as one of the main sources of fresh water, is affected by various man-made and natural factors, and has undergone different degrees of evolution. Groundwater is generally not renewable, so it is very important for rational exploitation and use of groundwater. This paper starts with the time series data of water level between groundwater and ground plane and makes prediction by deep learning method. RNN, Prophet and LSTM were used as the comparative experiment. The experiment showed that LSTM had more accurate prediction effect on the evaluation indexes MAE and RMSE.

Keywords

Neural Networks,LSTM, Groundwater level.

1. Introduction

The characteristics of groundwater are: the water temperature changes little in a year, usually 1°C-2°C higher than the annual average temperature; it is relatively clean, contains less organic matter, suspended solids and colloidal substances, and contains less microorganisms, and does not contain aquatic animals and Plants; the hardness of water is higher than that of river water, and the content of soluble salts is higher. Coupled with the non-renewable characteristics in general, it is destined to be a relatively scarce resource. With the expansion of population and the development of industry and agriculture, the shortage of water resources is becoming more and more serious, and people place more hope on groundwater. However, with the deepening of human industrialization and urbanization, the demand for groundwater continues to increase. But if uncontrolled over-extraction of groundwater results in a variety of man-made disasters[1]. Such as land subsidence, seawater intrusion, soil salinization, deterioration of groundwater quality, etc. Therefore, more scientific and sustainable mining strategies are needed[2].

2. Related Work

2.1. Neural Networks

Traditional time series forecasting generally uses the traditional mathematical statistical model ARMA level and its variants, such as ARIMA[3]. These methods perform well in statistics and forecasting linear data, but when predicting nonlinear data, such as discrete values in the data set, The effect is poor. With the emergence of neural networks, nonlinear problems have been solved. Like the human brain, various neural network structures can be formed to solve problems in different application scenarios[4]. However, with the increasing complexity of the network structure, the number of hyperparameters will increase exponentially, resulting in difficulties in adjusting hyperparameters and slow model training.

ISSN: 2664-9640

2.2. RNN(Recurrent Neural Network)

What this paper studies is time series data, so it should be sensitive to historical data. Although the fully connected neural network can predict a thing, the input of the previous data and the input of the latter data are completely independent, which makes it impossible to deal with some data with sequence information. The emergence of RNN not only optimizes the shortcomings of ordinary neural networks, but also remains sensitive to historical sequence information. Starting from the second neuron, the output of each neuron is the input of the next neuron, so the structure enables it to record the complete historical information of the sequence data[5]. The structure is shown in Figure 1:

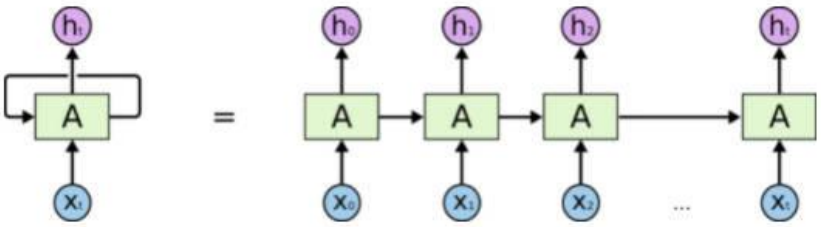

Figure 1: Structure diagram of RNN

On short-term sequence data, RNN can guarantee good prediction performance, but once the data set is too large or the historical information of sequence data is too much, RNN will have problems such as gradient disappearance or gradient explosion, resulting in difficulty in model training and serious errors in results.

2.3. Prophet

Time series forecasting has always been a difficult point in forecasting problems, and it is difficult for people to find a general model that is applicable to rich scenarios. This is because the background knowledge of each forecasting problem in reality, such as the data generation process, is often different, even if it is the same. The factors and degrees that affect these forecast values are often different. In addition, forecasting problems often require a lot of professional statistical knowledge, which brings difficulties to analysts, which makes the time series forecasting problem particularly complex. Traditional time series forecasting methods usually have the following defects:

Applicable time series data is too limited

Missing values need to be filled

Model lacks flexibility

weak guidance

Facebook open sourced the time series forecasting framework Prophet. Prophet is an open source library based on a decomposable (trend + season + holiday) model. Prophet fully integrates business background knowledge and statistical knowledge. It allows us to use simple and intuitive parameters for high-precision time series prediction, and supports automatic Define the effects of seasons and holidays. The official claim is to "allow ordinary people to draw professional conclusions like data analysts"[6]. It has been used in water conservancy[7]. The specific principles are as follows:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t \quad (1)$$

Among them, g(t) is the trend function, which models the non-periodic changes of time series data; s(t) represents cyclical changes (for example: weekly or annual seasonality, which can be changed according to actual business application scenarios) granularity); h(t) represents the effect of holidays; εt represents the error term. The specific process is shown in Figure 2:

ISSN: 2664-9640

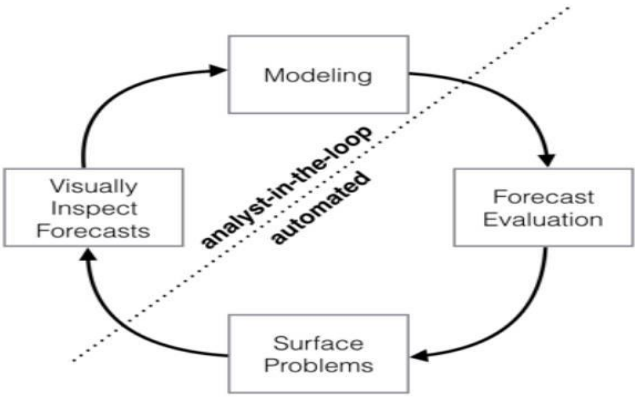

Figure 2: Structure diagram of FbProphet

2.4. LSTM(Long Short-Term Memory)

In order to solve the problem of RNN, the LSTM model algorithm adds a gate structure to solve the problem of gradient explosion and disappearance of RNN. It is a chain structure composed of repeated modules such as forgetting gates, inputs, output gates, memory units and activation functions[8]. Models have been applied to river levels[9], energy, and more[10]. Its structure is shown in Figure 3.

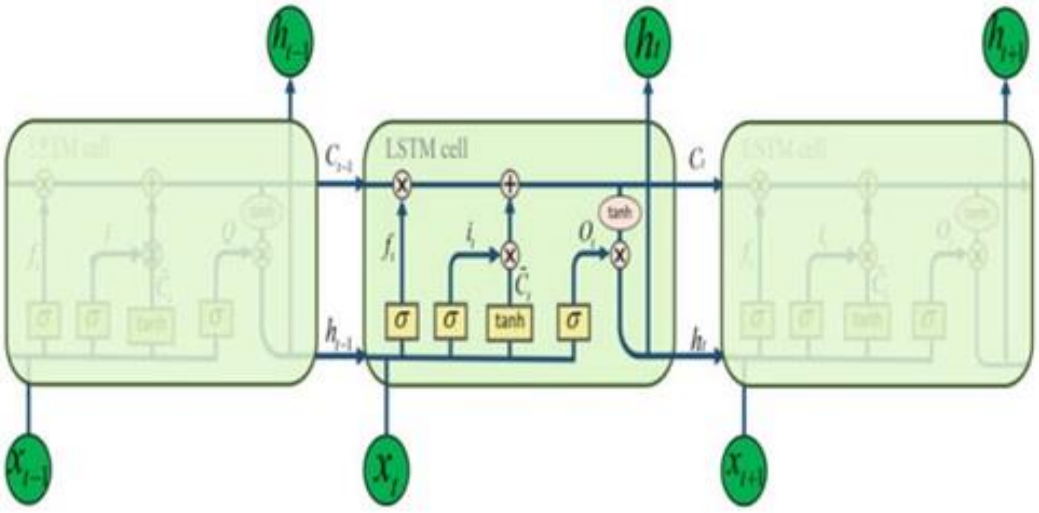

Figure 3: Structure diagram of LSTM

LSTM realizes the control of the information state of the memory cell through the forget gate and the input gate. The forget gate f_t decides which water level influence information of h_{t-1} and x_t to discard at time t; the input gate is used to determine how much water level influence information of x_t and h_{t-1} needs to be passed to c_t at time t in order to update the c_{t-1} storage Information. The LSTM uses the output gate to control the cell state c_t, and at the current time t, how much water level influence information in h_{t-1} and x_t is to be output. where each gate is as in formula 2:

Forget gate $$f_t = \sigma\left(W_f \cdot \left[h_{t-1}, x_t\right] + b_f\right)$$

Input gate $$i_t = \sigma\left(W_i \cdot \left[h_{t-1}, x_t\right] + b_i\right)$$

ISSN: 2664-9640

Output gate $$o_t = \sigma\left(W_o \cdot \left[h_{t-1}, x_t\right] + b_o\right)$$

Candidate cell state $$\tilde{c}_t = \tanh\left(W_c \cdot \left[h_{t-1}, x_t\right] + b_c\right) \quad (1)$$

Cell state $$c_t = f_t \otimes c_{t-1} + i_t \otimes \tilde{c}_t$$

Hidden state $$h_t = o_t \otimes \tanh\left(c_t\right)$$

3. Experiment and Analysis

3.1. Experiment procedure

3.1.1. Data sources and background

The dataset comes from (https://www.kaggle.com) and is a time series dataset of groundwater. In order to make more scientific and effective water resources (including above-ground water and groundwater), data analysis was performed on the data set collected from January 1, 2009 to June 30, 2020 in this area. The target feature is the groundwater table, the following is an explanation of the important features:

Rainfall: Rainfall (mm)

Temperature: temperature (°C)

Volume: This indicator represents the volume of water (m^3) extracted from the drinking water treatment plant

Hydrometry: Indicates groundwater level (m)

3.1.2. Experimental environment and methods

The model built by the python3.7-based deep learning library tensorflow. The configuration of the experimental computer is AMD Ryzen 4800H CPU + Nvidia GeForce GTX1650. And set parameter hyperparameters according to different models. This paper adopts the experimental method of comparison experiment, and the comparison model is ARIMA, RNN, PROPHET.

3.1.3. Data preprocessing

The division ratio of the training set and the test set is 8:2. The data set used contains very few missing values. The linear interpolation method is used to fill in, and the time feature is converted into an index, and then the person coefficient is used to determine the autocorrelation of each feature. , the closer it is to 1, the stronger the correlation. The specific calculation formula is Formula 3:

$$\rho = \mathrm{corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_x \sigma_Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - x)^2}\sqrt{\sum_{i=1}^{n}\left(y_j - y\right)^2}} \quad (3)$$

Resulting in correlation Figure 4:

ISSN: 2664-9640

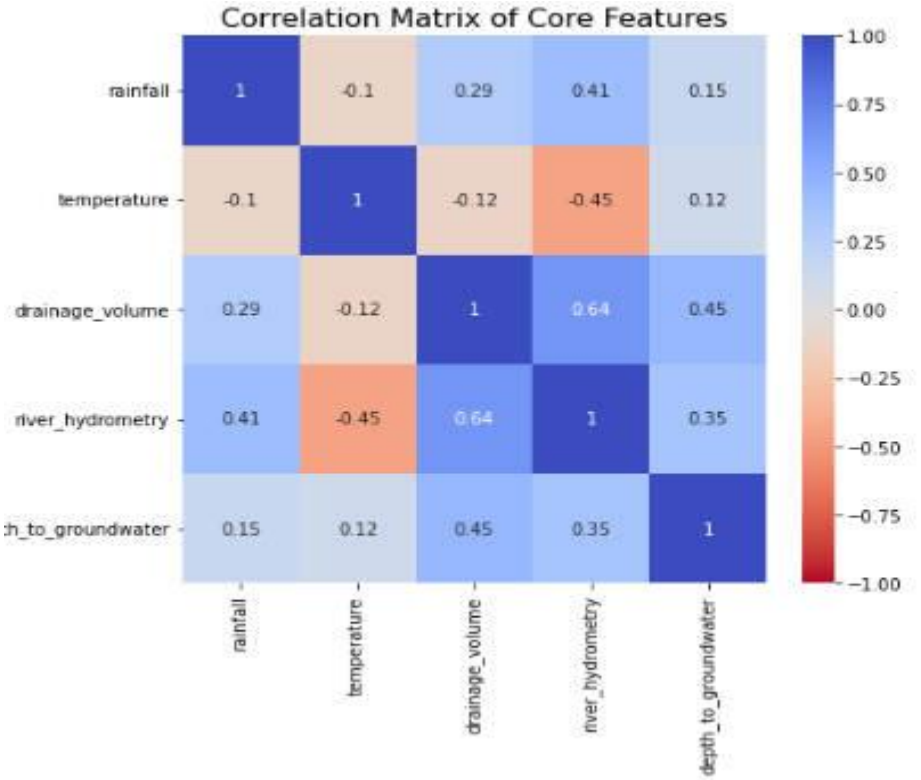

Figure 4: Correlation of individual features

The target feature visualization is shown in Figure 5:

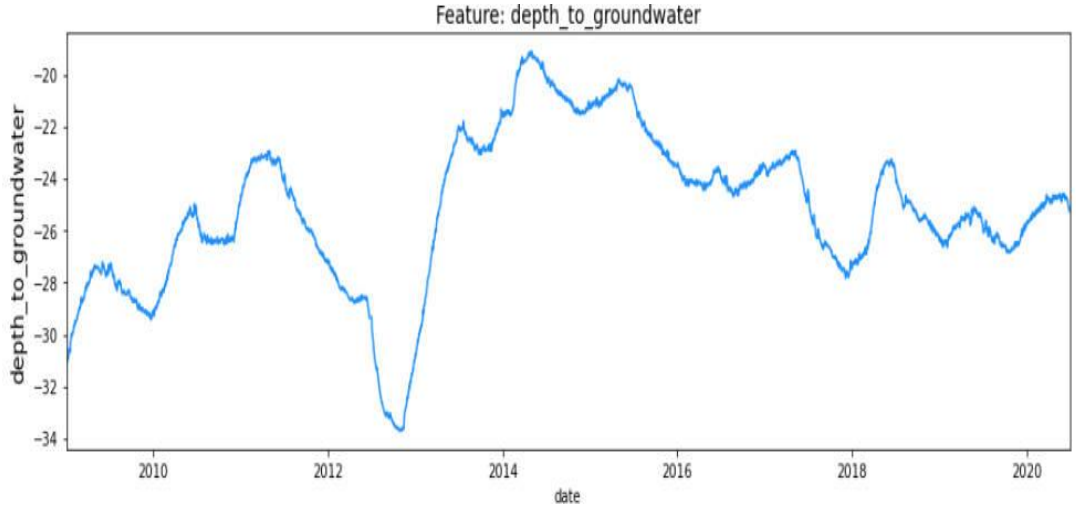

Fig 5: Visualization of target features

3.2. Evaluation indicators

In this paper, two evaluation indicators are used to evaluate the prediction effect of the model, namely Mean Absolute Error (MAE) and Root-Mean Square Error (RMSE). MAE is the squared difference between the predicted value and the true value. RMSE is sensitive to larger errors. The smaller the value of the two indicators, the better. The specific calculation of the indicator is shown in formula 4-5:

$$\mathrm{MAE} = \frac{1}{n}\sum_{n}^{i=1}\left|y_i - \hat{y}_i\right| \tag{4}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_t - \hat{y}_t\right)^2}{n}} \tag{5}$$

ISSN: 2664-9640

Among them, n in the formula is the total number of samples, y is the actual value, and $\hat{y}$ is the predicted value.

3.3. Result and Analysis

3.3.1. Result

After a series of experiments, the experimental results of each model are shown in Figure 6-9:

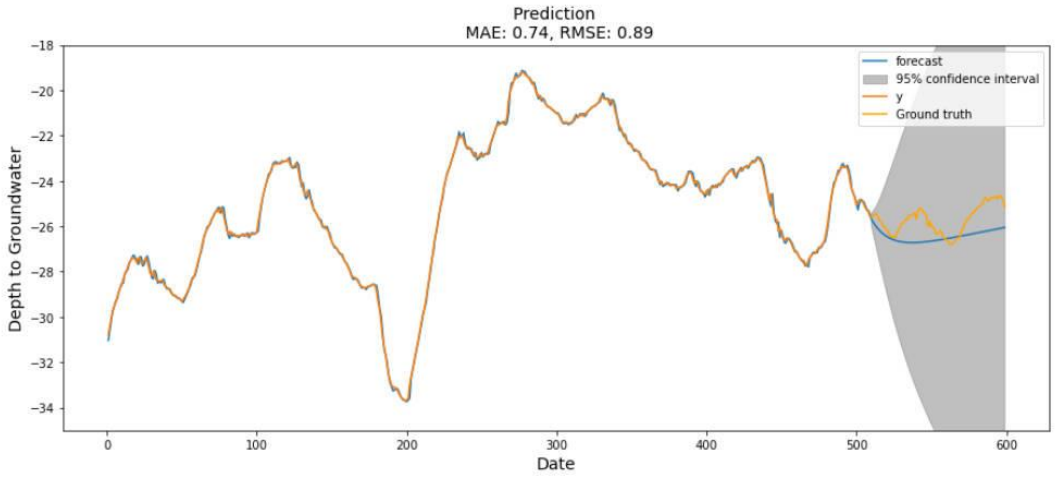

FIG 6: ARIMA's forecast renderings

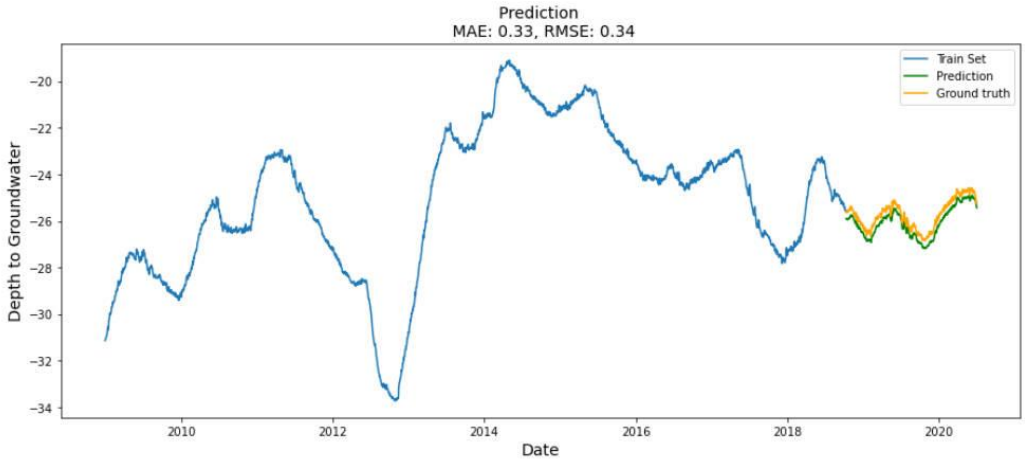

FIG 7: Prediction rendering of RNN

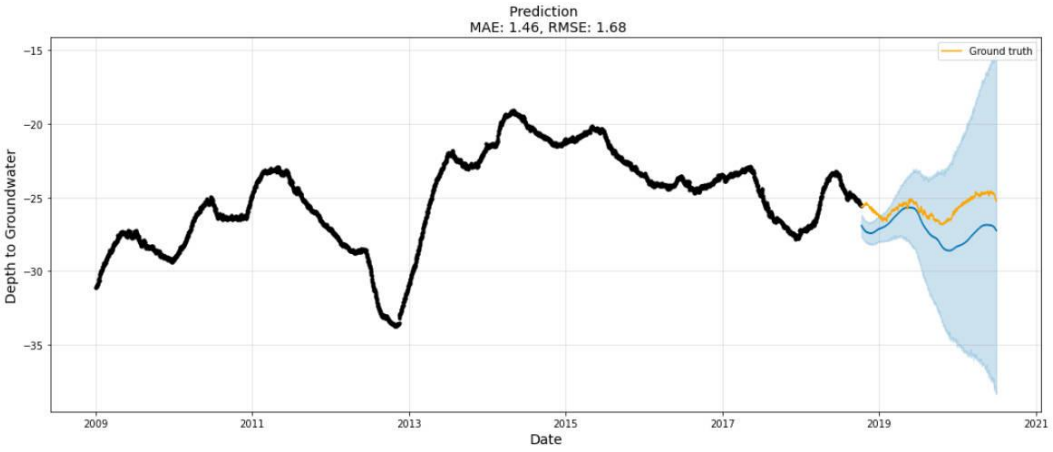

FIG 8: PROPHET's forecast renderings

ISSN: 2664-9640

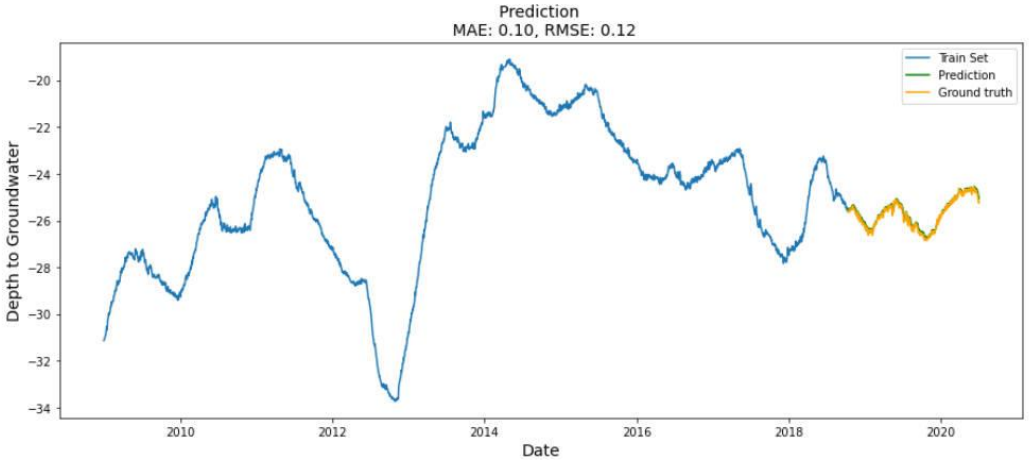

FIG 9: Prediction rendering of LSTM

3.3.2. Analysis

The statistical results of the evaluation indicators of each model are shown in Table 1.

Table 1: Model prediction result table

Model	ARIMA		RNN		PROPHET		LSTM	
Index	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Result	0.74	0.89	0.33	0.34	1.46	1.68	0.10	0.12

Judging from the above prediction effect in the test set, the prediction effect of the ARIMA model is not so bad, but its prediction effect looks too smooth and is not sensitive to discrete values or abnormal predictions. Prophet has a certain sensitivity to discrete and outliers, but the prediction effect is not ideal. The prediction of RNN is more accurate, but if the data time length is doubled, there will be unpredictable situations. LSTM solves the shortcomings of the above two models. For long-term water level time series, the gradient of RNN will not disappear or explode. And the prediction effect is more accurate.

4. Conclusion

Because the characteristics of groundwater determine its scarcity, but due to the increasing demand for groundwater due to human urbanization and industrialization, scientific and effective groundwater exploitation strategies or plans that can protect the natural environment are particularly important at this time. The groundwater level prediction in this paper has technical reference value for formulating future exploitation plans. Subsequent research will add more factors that affect groundwater into the model, such as large geological changes. Enables more accurate groundwater predictions.

References

[1] Li Shuai et al. The Recent Progress China Has Made in Green Mine Construction, Part I: Mining Groundwater Pollution and Sustainable Mining[J]. International Journal of Environmental Research and Public Health, 2022, 19(9) : 5673-5673.

[2] Narsimha Adimalla and Hui Qian and M.J. Nandan. Groundwater chemistry integrating the pollution index of groundwater and evaluation of potential human health risk: A case study from hard rock terrain of south India[J]. Ecotoxicology and Environmental Safety, 2020, 206 : 111217-111217.

[3] Singh S N, Mohapatra A. Repeated wavelet transform based ARIMA model for very short-term wind speed forecasting[J]. Renewable energy, 2019, 136: 758-768.

[4] Hadjisolomou Ekaterini et al. Modelling Freshwater Eutrophication with Limited Limnological Data Using Artificial Neural Networks[J]. Water, 2021, 13(11) : 1590-1590.

ISSN: 2664-9640

[5] Zhang Xiaohan et al. Episodic memory governs choices: An RNN-based reinforcement learning model for decision-making task[J]. Neural Networks, 2021, 134 : 1-10.

[6] Taylor S J, Letham B. Forecasting at scale[J]. The American Statistician, 2018, 72(1): 37-45.

[7] De Filippis Tiziana et al. Hydrological Web Services for Operational Flood Risk Monitoring and Forecasting at Local Scale in Niger[J]. ISPRS International Journal of Geo-Information, 2022, 11(4) : 236-236.

[8] Noor Fahima et al. Water Level Forecasting Using Spatiotemporal Attention-Based Long Short-Term Memory Network[J]. Water, 2022, 14(4) : 612-612.

[9] Liu Yu et al. Short Term Real-Time Rolling Forecast of Urban River Water Levels Based on LSTM: A Case Study in Fuzhou City, China[J]. International Journal of Environmental Research and Public Health, 2021, 18(17) : 9287-9287.

[10] El Bourakadi Dounia and Yahyaouy Ali and Boumhidi Jaouad. Intelligent Energy Management for Micro-Grid based on Deep Learning LSTM prediction Model and Fuzzy Decision-Making[J]. Sustainable Computing: Informatics and Systems, 2022, : 100709-.